Assignment: Advanced Regression
Subjective Questions

Question 1
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

The optimal value for the Ridge regression model = 10
The optimal value for the Lasso regression model = 0.01

The changes to the model on doubling the alpha values were not very significant. Regularization did not have much of an impact since the model was fairly generalized.

In the ridge regression model, the R2_score dropped by 0.00082.
In the lasso regression model, the R2_score dropped by 0.008

The most important variable remains unchanged after doubling the alpha value.
The most important variable in ridge regression is '2ndFlrSF'
The most important variable in lasso regression is 'OverallQual'

Question 2
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

I would choose the Ridge regression model. This model had a better r2_score compared to the lasso model. In fact, the ridge model performed significantly better on the test set that the training set

Question 3

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

On creating the lasso regression model, the top 5 variables were:

- OverallQual
- 2ndFlrSF
- 1stFlrSF
- YearBuilt
- GarageCars

After dropping these features, the top 5 predictors were

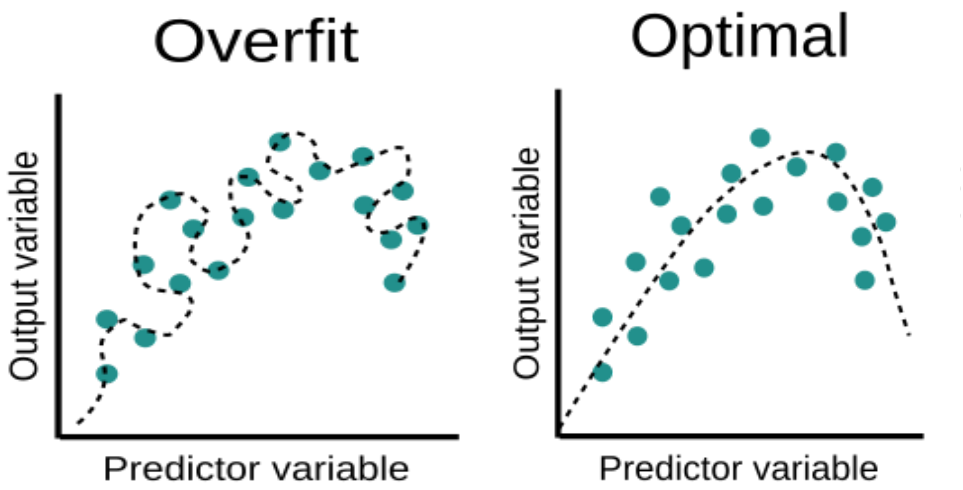
- ExterQual_Gd
- FireplaceQu_No_fireplace
- Neighborhood_NridgHt
- Neighborhood_NoRidge
- BsmtExposure_Gd

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

In order for a model to be robust and generalisable, it must learn the pattern of the data, and not the data itself.

One of the major issues that plaque machine learning is overfitting. Overfitting is when the model learns individual data points instead of the trend of the data.



One of the main reasons overfitting occurs is a high model complexity.
There are 2 main factors that lead to a complex model.

1. The magnitude of the coefficients
2. The number of features

In order to keep the model simple and less complex, we incorporate various techniques such as 'recursive feature elimination' to get rid of variables that are correlated to each other and hence do not have much significance.
The other is regularization. This reduces the magnitude of the coefficients towards zero.

The implications of a complex model are that it yields a very high accuracy score on the training set since it learns all the data points.
But it fails to perform on a data set that it has not been trained on. It yields a low accuracy on the test set.

A robust and generalized model that has learnt the trend of the data performs fairly well on the training set and is able to generalize and perform well on data that it has not seen before